Run Descriptions: TREC 2009 Legal Track, Interactive Task

The descriptions of the experimental runs are taken from the form information provided at submission time. Fields which were left blank in the form are omitted below. Please consult the participant and track overview papers for more information about the experiments conducted.

Run ADI2009Topic204 (Applied Discovery)
Task: interactive
Document versions used: native
Time spent with topic authority per topic (hours):
Topic 204: 6
Description of this run:
Topic 204 / we used ORA e-discovery tool with FAST search; key term and boolean searches to pull results; document sampling to test and refine. Results file contains 80,114 document IDs.

Run Clearwell09i (Clearwell Systems Inc.)
Task: interactive
Document versions used: native, text
Time spent with topic authority per topic (hours):
Topic 201: 4
Topic 202: 6
Description of this run:
Clearwell E-Discovery platform V4.5 is used for the execution of the legal track interactive task topic 201 and topic 202.

Run clearwell01 (Clearwell Systems Inc.)
Task: interactive
Document versions used: native
Time spent with topic authority per topic (hours):
Topic 205: 8
Description of this run:
Clearwell E-Discovery platform v4.5 is used for topic 205 request production.

Run CGSHBCK (Cleary Gottlieb Steen & Hamilton with Backstop LLP)
Task: interactive
Document versions used: text
Description of this run:
Run from Cleary Gottlieb Steen and Hamilton - Backstop team

Run CGSHBCK1 (Cleary Gottlieb Steen & Hamilton with Backstop LLP)
Task: interactive
Document versions used: text
Description of this run:

Replacement run from Cleary Gottlieb Steen and Hamilton - Backstop team. This is meant to substitute for the earlier run.

Run CGSHBCK2 (Cleary Gottlieb Steen & Hamilton with Backstop LLP)
Task: interactive
Document versions used: text
Description of this run:

Replacement run from Cleary Gottlieb Steen and Hamilton - Backstop team. This is meant to substitute for the earlier run.

Run Equivio205R1 (Equivio)
Task: interactive
Document versions used: text
Time spent with topic authority per topic (hours):
Topic 205: 11.5
Description of this run:

The Equivio run used Equivio>Relevance, an expert-guided system for assessing document relevance. The system feeds statistically selected samples of documents to an expert (an attorney familiar with the case), who marks each sample as relevant or not. The expert's decisions are used to train the software to estimate document relevance. Using a statistical model to determine when the software training process has optimized, the system then calculates graduated relevance scores for each document in the collection.

Run Equivio207R1 (Equivio)
Task: interactive
Document versions used: text
Time spent with topic authority per topic (hours):
Topic 207: 5
Description of this run:

The Equivio run used Equivio>Relevance, an expert-guided system for assessing document relevance. The system feeds statistically selected samples of documents to an expert (an attorney familiar with the case), who marks each sample as relevant or not. The expert's decisions are used to train the software to estimate document relevance. Using a statistical model to determine when the software training process has optimized, the system then calculates graduated relevance scores for each document in the collection.

Run H52009 (H5)
Task: interactive
Document versions used: text
Time spent with topic authority per topic (hours):
Topic 204: 8.5
Description of this run:

H5 has submitted an assessment of documents our system has identified as responsive to topic 204 of TREC's Legal Track Interactive Task. Our system identified 2994 documents as responsive to topic 204. The H5 system combined human expertise and advanced search and information retrieval technologies to assess the totality of the corpus under investigation.

Run IntegreonB (Integreon)
Task: interactive
Document versions used: native
Time spent on each topic (hours):
Topic 205: 40
Time spent with topic authority per topic (hours):
Topic 205: 4
Description of this run:

We considered the entire message unit responsive if any part of the unit was responsive. We have also listed all items from all responsive message units, all emails and attachments from each "family". We are submitting for Topic 205, only.

Run LogikIT09t (Logik Systems, Inc)
Task: interactive
Document versions used: text
Time spent with topic authority per topic (hours):
Topic 206: 1
Topic 207: 1
Description of this run:

Documents were classified with a Naive Bayes classifier which was trained from a set of internally tagged documents.

Run pittsis09 (School of Info Sciences, Univ of Pittsburgh)
Task: interactive
Document versions used: text
Time spent with topic authority per topic (hours):
Topic 201: 1
Description of this run:

We designed an experiment to investigate into the information seeking behavior of users when conducting e-discovery task. Our focus is on the collaboration among searchers. We observed an expert with legal background and an information retrieval expert working collaboratively on topic 201. How they collaborate with each other to complete the task and what the characteristics of the collaborative information behavior (CIB) are.

Run buffalo (Univ at Buffalo, State Univ of New York)
Task: interactive
Document versions used: text
Time spent on each topic (hours):
Topic 203: 50
Time spent with topic authority per topic (hours):
Topic 203: 0.5
Description of this run:

We combined results of about 15 very specific queries with results of one generic query

Run watlint (University of Waterloo)
Task: interactive
Document versions used: native, text
Time spent on each topic (hours):
Topic 201: 15
Topic 202: 25
Topic 203: 10
Topic 207: 30
Time spent with topic authority per topic (hours):
Topic 201: 2
Topic 202: 1.5
Topic 203: 1
Topic 207: 3.1
Predicted precision/recall per topic:
Topic 201: 0.7 / 0.9
Topic 202: 0.7 / 0.9
Topic 203: 0.3 / 0.5
Topic 207: 0.9 / 0.95
Description of this run:

Interactive search and judging followed by active machine learning with human reviewer in loop. Recall estimated by fitting censored normal distribution to machine learning scores for responsive documents, factoring in an estimate of 90% inter-assessor agreement on non-relevant documents (derived from past experience). Precision estimated assuming 70% as an upper bound on human agreement (based on past experience), reduced for topic 203 due to poor fit of the learning model, coupled with uncertainty in review. Topic 207 estimates are confounded (upwards) by the fact that about 2/3 of the responsive documents are vacuous ".URL" attachments, which were handled as a special case.

Run CompCustIT09 (ZL Technologies, Inc.)
Task: interactive
Document versions used: native, text
Time spent with topic authority per topic (hours):
Topic 203: 1
Description of this run:

In this submission, the emails were reduplicated to approximately 3 million emails and distributed across 104 custodian mailboxes. The reduplicated emails combined the text extraction for the message body with the native files for the attachments to create IETF RFC-2822 MIME emails. The custodians were further associated with titles and department information. The team identified and prioritized the most likely custodians with relevant information based on the titles and department. Once the custodians had been identified, all their email was pulled and made available for review. A couple of other custodians were identified through a enterprise subject-based search without using title and department to see if the custodian identification method could miss important custodians with substantial volumes of relevant email. A variety of search and analytics techniques were used in conjunction with guidance from the Topic Authority. In this submission, the top 4 prioritized custodians based on title and department are included. The email's use an ID we term the "TREC Parent ID" which is the DocId after removing TREC identified duplicates. This ID is the broadest ID used in this study. The DocID is termed the "TREC ID." Since the DocID is not unique in a fully reduplicated set, a further ID, the "JWID" was created as a unique identifier to assist in performing the analysis. Although the top 4 custodians are presented, more custodians were analyzed and additional information can be submitted. 4 out of 104 possible custodians is 3.8% of the user population while in the overall set, the 104 custodians represents 0.48% of the approximately 22,000 Enron employees.

Run CompEntrIT09 (ZL Technologies, Inc.)
Task: interactive
Document versions used: native, text
Time spent with topic authority per topic (hours):
Topic 203: 1
Description of this run:

In this submission, the emails were reduplicated to approximately 3 million emails and distributed across 104 custodian mailboxes. The reduplicated emails combined the text extraction for the message body with the native files for the attachments to create IETF RFC-2822 MIME emails. This created a scenario similar to a eDiscovery scenario before processing. The emails were ingested into the ZL eDiscovery review platform where a variety of search and analytics capabilities were applied to the data including full text search, wildcard search, auto-classification, concept search, faceted search etc. These techniques were used in conjunction with interactive guidance from the Topic Authority.